Regresión Lineal

Universidad Tecnológica Nacional
Facultad Regional Paraná

Regresión Lineal

Objetivos

- Ajustar una RL a un par de variables numéricas.
- Valorar el ajuste global del modelo mediante ANOVA.
- Realizar inferencia sobre los parámetros del modelo.
- Evaluar el modelo mediante técnicas diagnósticas.
- Interpretar los valores de los parámetros de la recta y sus intervalos de confianza.
- Realizar predicciones y sus intervalos de confianza.
- Graficar los datos y la recta de RL ajustada.
- Validación cruzada.

Regresión Lineal Simple

El modelo de regresión lineal simple describe la relación entre dos variables X e Y de la siguiente manera:

$$Y = \beta_0 + \beta_1 * X_1 + \epsilon$$

- Coeficientes β_i (e.g. β_0 es el intercepto)
- Errores o residuos.

- ✓ Estimación de la RLS
- ✓ Coeficiente de determinación
- ✓ Prueba de significación para RLS
- ✓ Intervalos de confianza y de predicción para RLS
- ✓ Diagnóstico para RLS

Ejemplo: datos *Duncan*

Prestigio y otras características de 45 tipos de trabajos en U.S. en 1950. 45 filas (casos) y 4 columnas (variables):

- ✓ " type" (tipo de ocupación: prof -profesionales y directivos-, wc -de cuello blanco-, bc -de cuello azul-).
- ✓ "income" (% de los varones que ganan $3500 ó más en 1950).
- ✓ "education" (% de los varones graduados de secundaria).
- ✓ "prestige" (% de los evaluadores en NORC que calificaron la ocupación como excelente o bueno en prestigio).

```
library(car)
data(Duncan)
```

Ejemplo: datos *Duncan*

Vamos a intentar explicar el nivel de ingresos Y="income" (variable dependiente, VD) a partir del nivel educativo X="education" (variable independiente, VI).

```
head(Duncan)
```

```
##            type income education prestige
## accountant prof     62        86       82
## pilot      prof     72        76       83
## architect  prof     75        92       90
## author     prof     55        90       76
## chemist    prof     64        86       90
## minister   prof     21        84       87
```

Ejemplo: datos *Duncan*

```
fitLS <- lm(income ~ education, data = Duncan)
summary(fitLS)
```

```
Call:
lm(formula = income ~ education, data = Duncan)
```

Fórmula

```
Residuals:
    Min      1Q  Median      3Q     Max
-39.572 -11.346  -1.501   9.669  53.740
```

Diagnóstico

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.6035     5.1983   2.040   0.0475 *
education     0.5949     0.0863   6.893 1.84e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Coeficientes

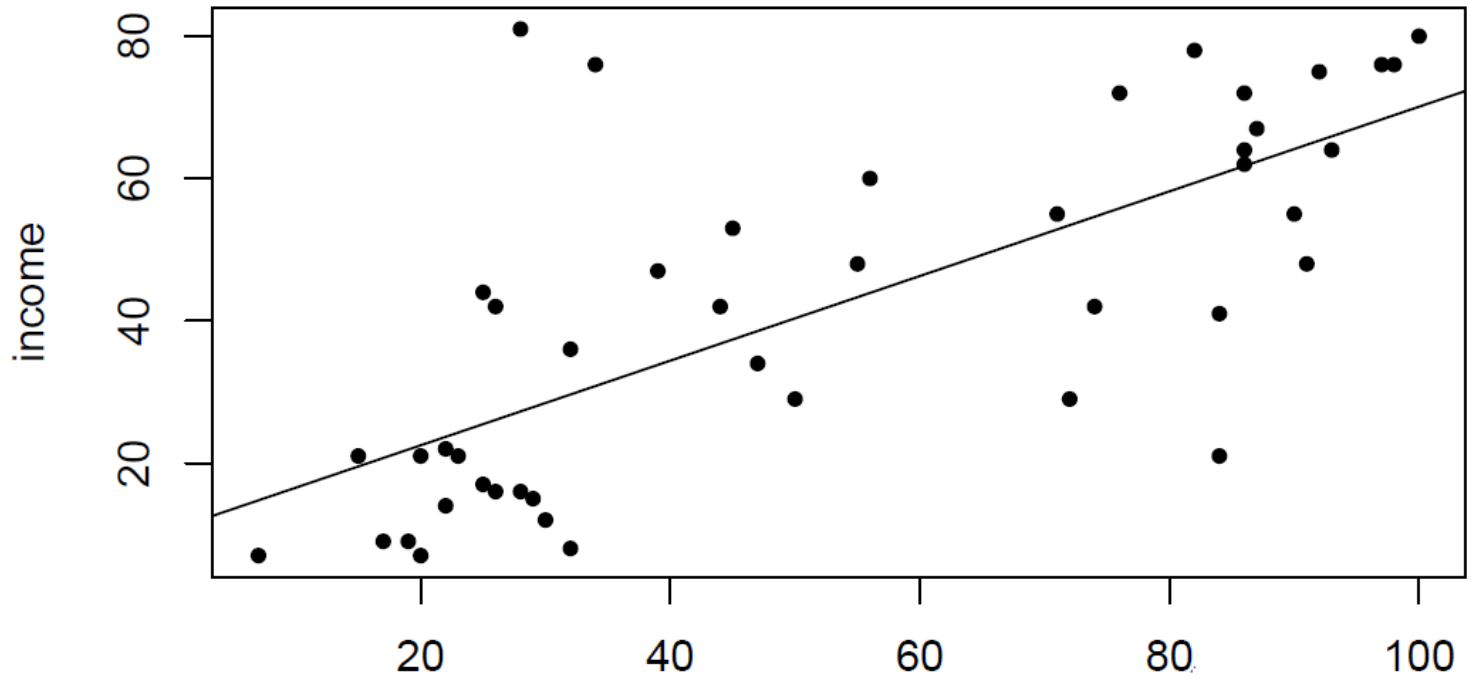
```
Residual standard error: 17.04 on 43 degrees of freedom
Multiple R-squared:  0.5249,    Adjusted R-squared:  0.5139
F-statistic: 47.51 on 1 and 43 DF,  p-value: 1.84e-08
```

Bondad de ajuste

Ejemplo: datos *Duncan*

Representación gráfica con la recta ajustada y los datos.

```
plot(income~education, data=Duncan, pch=20)
abline(fitLS)
```



Ejemplo: datos *Duncan*

Ajuste global del modelo (ANOVA).

Partición de varianza: qué parte de la variabilidad de la respuesta es explicada por su relación con las variables predictoras y qué parte no es explicada por dicha relación (residual).

Esto permitirá contrastar si el modelo es significativo o no.

En nuestro ejemplo:

$$F(1,43)=47.51,\ p<0.001$$

Ejemplo: datos *Duncan*

Parámetros del modelo.

Estimaciones de los parámetros β del modelo.

Contrastes de hipótesis sobre cada parámetro ($H_0: \beta_i = 0$).

Prueba *t de Student* en cada parámetro.

En nuestro ejemplo:

- intercepto $\beta_0 = 10.6035$, «casi» no significativo ($p = 0.0475$)

- pendiente $\beta_1 = 0{,}5949$, significativo ($p < 0.001$)

Ejemplo: datos *Duncan*

Bondad de ajuste R^2.

Medida de la eficacia del modelo de regresión: coeficiente de determinación R^2.

- ✓ Toma valores en el intervalo 0-1.
- ✓ Mide el porcentaje de variabilidad en los datos que viene explicada por el modelo, por lo que un valor cercano a 1 significa que el modelo es bastante efectivo.

NOTA: Al agregar más variables al modelo el R^2 aumenta, por lo cual en modelos de regresión múltiple se aconseja utilizar el R^2 ajustado, que ajusta su valor para dar cuenta del número de variables incluidas en el modelo.

En nuestro ejemplo es bajo $R^2 = 0.525$.

Ejemplo: datos *Duncan*

Supuestos del modelo:

1. **Linealidad**. La relación es constante. Gráfico de dispersión.

2. **Independencia de los residuos**. Los valores y_i y el término de error *i* son independientes entre sí. Prueba de Durbin-Watson.

3. **Normalidad de los residuos**. Implica que los valores y_i y el término de error *i* tienen distribución normal para cada valor de x_i . Prueba de Kolmogorff-Smirnov o Shapiro-Wilks; métodos gráficos (gráficos de normalidad de tipo QQ cuantiles o PP proporciones, e histograma).

4. **Homogeneidad de varianza**. Los valores y_i y el término de error *i* tienen la misma varianza para cada x_i .

Ejemplo: datos *Duncan*

Observaciones atípicas, extrañas o influyentes. -*outliers* –

- ✓ Atípicas con respecto al eje de abscisas *X*.
- ✓ Atípicas en relación de eje de ordenadas *Y* .
- ✓ Atípicas respecto tanto a las abscisas como a las ordenadas.

Para detectarlos evaluamos los residuos respecto al: apalancamiento (*leverage*), distancia de Cook, DFFITS o DFBETAS.

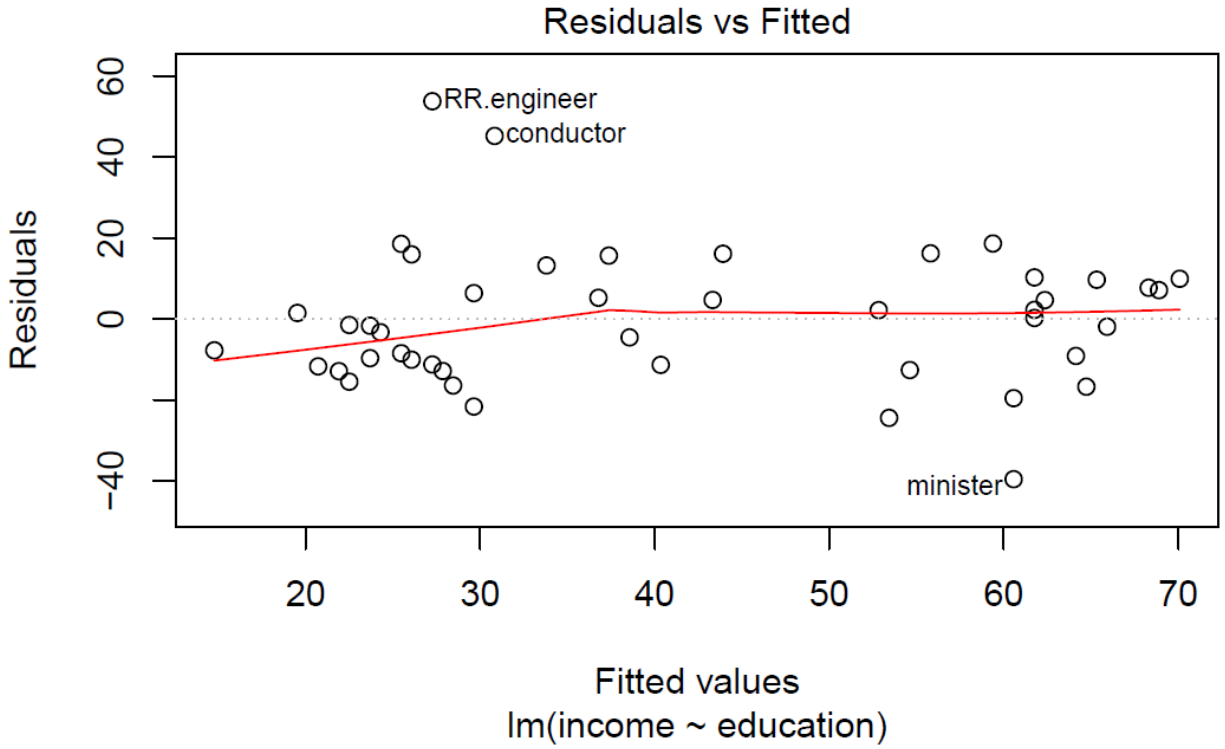
Ejemplo: datos *Duncan*

Diagnóstico:

```
par(mfrow=c(2,2))
plot(fitLS)
```

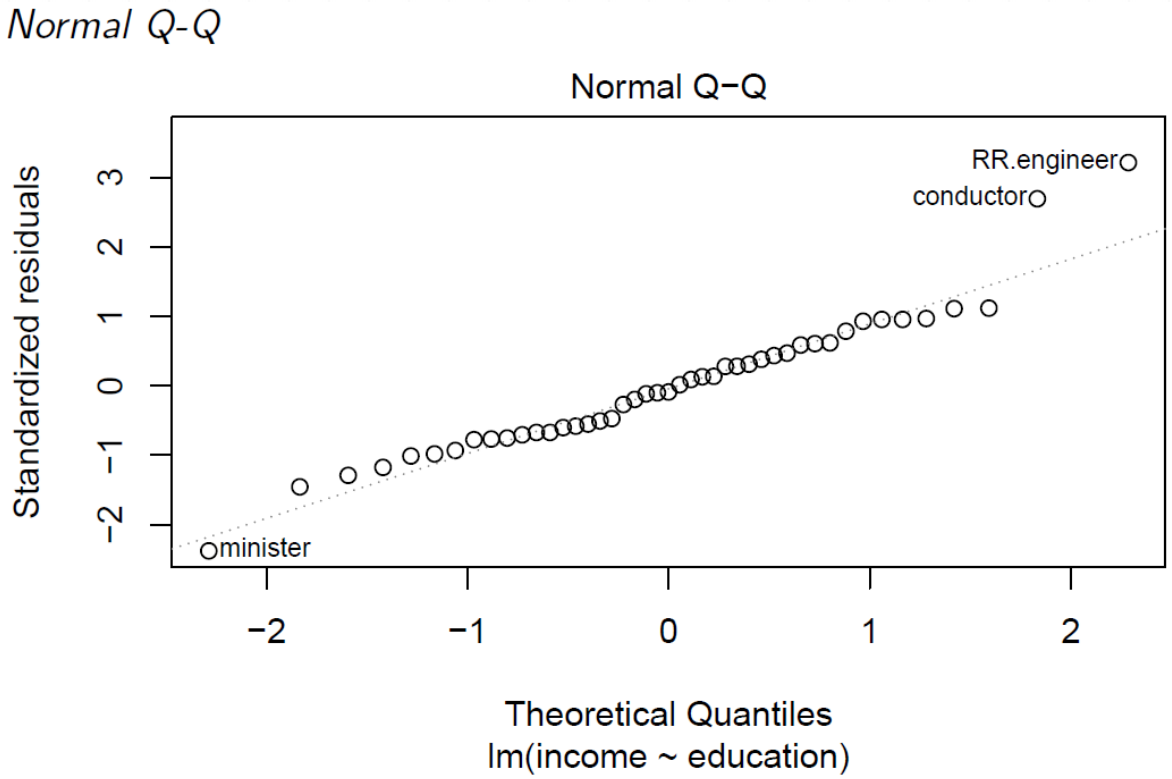
Ejemplo: datos *Duncan*

Residual vs Fitted



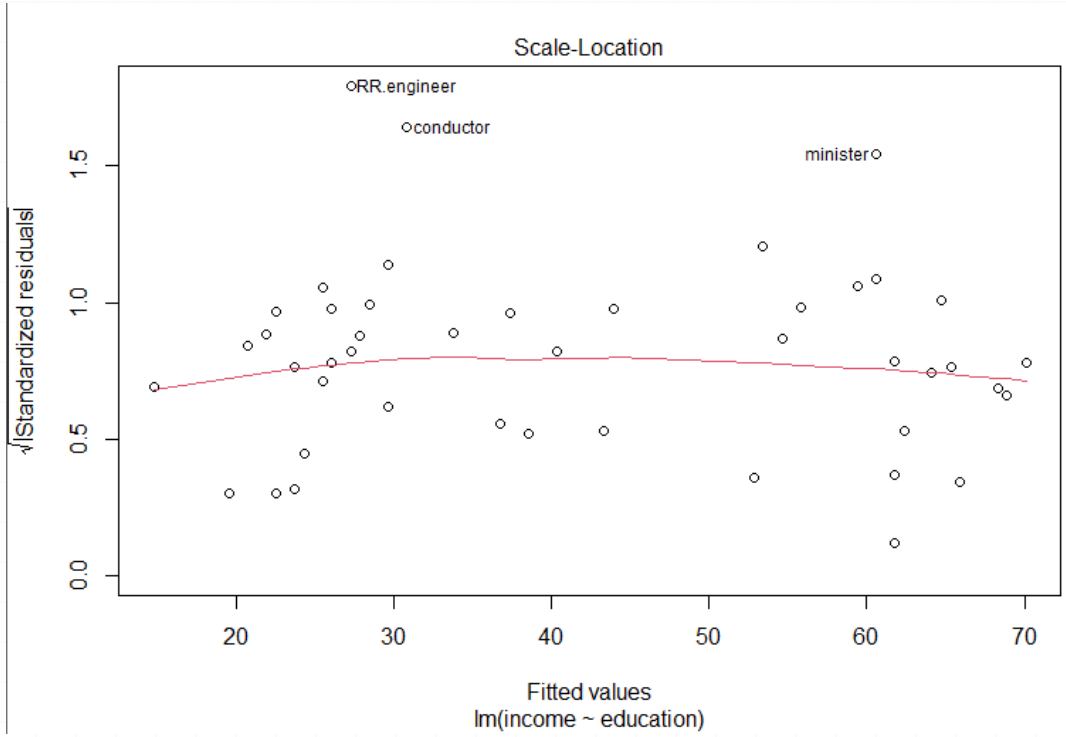
Media cero, no hay heterocesdasticidad y ni problemas de linealidad.
Residuos no correlacionados.

Ejemplo: datos *Duncan*



No hay evidencias claras de heterocesdasticidad.
Muestra la existencia de 3 datos atípicos.

Ejemplo: datos *Duncan*



Muestra la existencia de 3 datos atípicos.

Ejemplo: datos *Duncan*

Predicciones:

```
# Intervalos de confianza para la respuesta media
(res.pred1 <- predict(fitLS,
                      list(education= c(10, 50, 90)),
                      interval="confidence"))
```

```
##        fit       lwr      upr
## 1 16.55209  7.547121 25.55706
## 2 40.34647 35.205386 45.48755
## 3 64.14085 55.852073 72.42962
```

Ejemplo: datos *Duncan*

Predicciones:

```
# Intervalos de predicción
(res.pred2 <- predict(fitLS,
                      list(education= c(10, 50, 90)),
                      interval="prediction"))
```

```
##        fit       lwr      upr
## 1 16.55209 -18.96651 52.07069
## 2 40.34647   5.60583 75.08711
## 3 64.14085  28.79703 99.48466
```

Ejemplo: datos *Duncan*

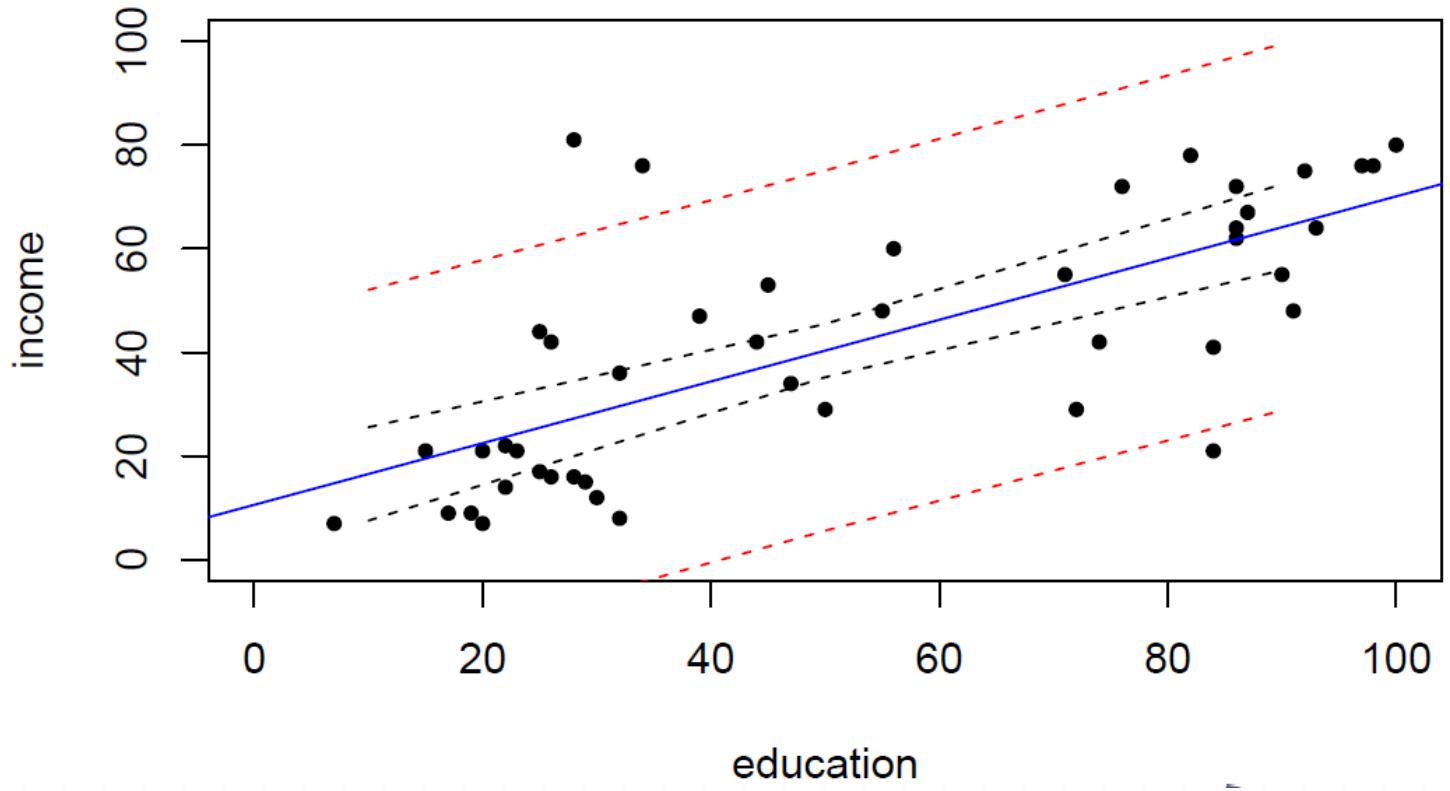
Representación del modelo:

```
# aumentamos los límites del gráfico
plot(income~education, data=Duncan, xlim=c(0,100),
     ylim=c(0,100), pch=20)
abline(fitLS,col="blue")

lines(c(10, 50, 90), res.pred1[, 2], lty = 2)
lines(c(10, 50, 90), res.pred1[, 3], lty = 2)
lines(c(10, 50, 90), res.pred2[, 2], lty = 2, col = "red")
lines(c(10, 50, 90), res.pred2[, 3], lty = 2, col = "red")
```

Ejemplo: datos *Duncan*

Representación del modelo:



Ejemplo: datos *Duncan*

Validación cruzada:

```
library(caret)
train(income~education, data=Duncan, method="lm")
```

```
Linear Regression

45 samples
 1 predictor

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 45, 45, 45, 45, 45, 45, ...
Resampling results:

  RMSE      Rsquared   MAE
  17.59711  0.5129526  13.86149

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Obtenemos que el ajuste es bueno.

Ejemplo: datos *Duncan*

Resumen:

- ✓ Explicamos un bajo porcentaje de variabilidad (52.5%), por lo que parece necesario un modelo de regresión múltiple.

- ✓ Existen *outliers* que influyen en la recta, por lo que los modelos robustos parecen ser más adecuados en este caso.